BIG DATA ON TRIAL: RESEARCHING SYNTACTIC ALTERNATIONS IN GLOWBE AND ICE

D2E Conference, Helsinki

October 21, 2015

Benedikt Heller
benedikt.heller@kuleuven.be

Melanie Röthlisberger
melanie.rothlisberger@kuleuven.be

KU LEUVEN

OVERVIEW

1. Introduction
2. Data
3. Methodology
4. Results
5. Summary & Conclusion

INTRODUCTION

PROJECT

- "Exploring probabilistic grammar(s) in varieties of English around the world"
- 5-year project (2013–2018), KU Leuven
- Three alternations
  - Particle placement: Jason Grafmiller
  - Dative alternation: Melanie Röthlisberger
  - Genitive alternation: Benedikt Heller
- Principal investigator: Benedikt Szmrecsanyi
- First results: Szmrecsanyi et al. (2016)

THE “ENGLISH WORLD-WIDE PARADIGM”

- Wide range of postcolonial varieties
- topics: scope, limits, parameters of variation; extent to which structural make-up of varieties of E can be predicted by communicative needs of colonizers/colonized

  (e.g. Kachru 1992; Schneider 2007; Mesthrie and Bhatt 2008)
- shortcoming: an often primarily descriptive interest in the variable presence/absence of features, or in usage frequencies of features

THE PROBABILISTIC GRAMMAR FRAMEWORK

- rely on the variation-centered, usage- and experience-based probabilistic grammar framework developed by Joan Bresnan and collaborators

  (e.g. Bresnan 2007; Bresnan and Ford 2010a; Wolk et al. 2013)

  1. syntactic variation – and change – is **subtle, gradient & probabilistic** rather than categorical in nature

     (Labov 1982; Bresnan and Hay 2008a)
  2. linguistic knowledge includes **knowledge of probabilities**, and speakers have powerful predictive capacities

     (Gahl and Garnsey 2004; Gahl and Yu 2006)

RESEARCH QUESTION

Desire to complement ICE data with data from GloWbE

RESEARCH QUESTION

Desire to complement ICE data with data from GloWbE

- Increasingly central part of the language, so it is crucial that we look at web-data more
- There are critics, so it's important to see how ICE and GloWbE differ

RESEARCH QUESTION

Desire to complement ICE data with data from GloWbE

- Increasingly central part of the language, so it is crucial that we look at web-data more
- There are critics, so it's important to see how ICE and GloWbE differ

How similar are syntactic alternation patterns in GloWbE to the patterns in ICE?

DATIVE ALTERNATION

(1) He gives [Mary]$_{\text{recipient}}$ [a present]$_{\text{theme}}$

ditransitive dative

(2) He gives [a present]$_{\text{theme}}$ to [Mary]$_{\text{recipient}}$

prepositional dative

GENITIVE ALTERNATION

(3) [The president]$_{possessor}$'s [speech]$_{possessum}$

***s*-genitive**

(4) The [speech]$_{possessum}$ of [the president]$_{possessor}$

***of*-genitive**

CHOICE CONTEXT

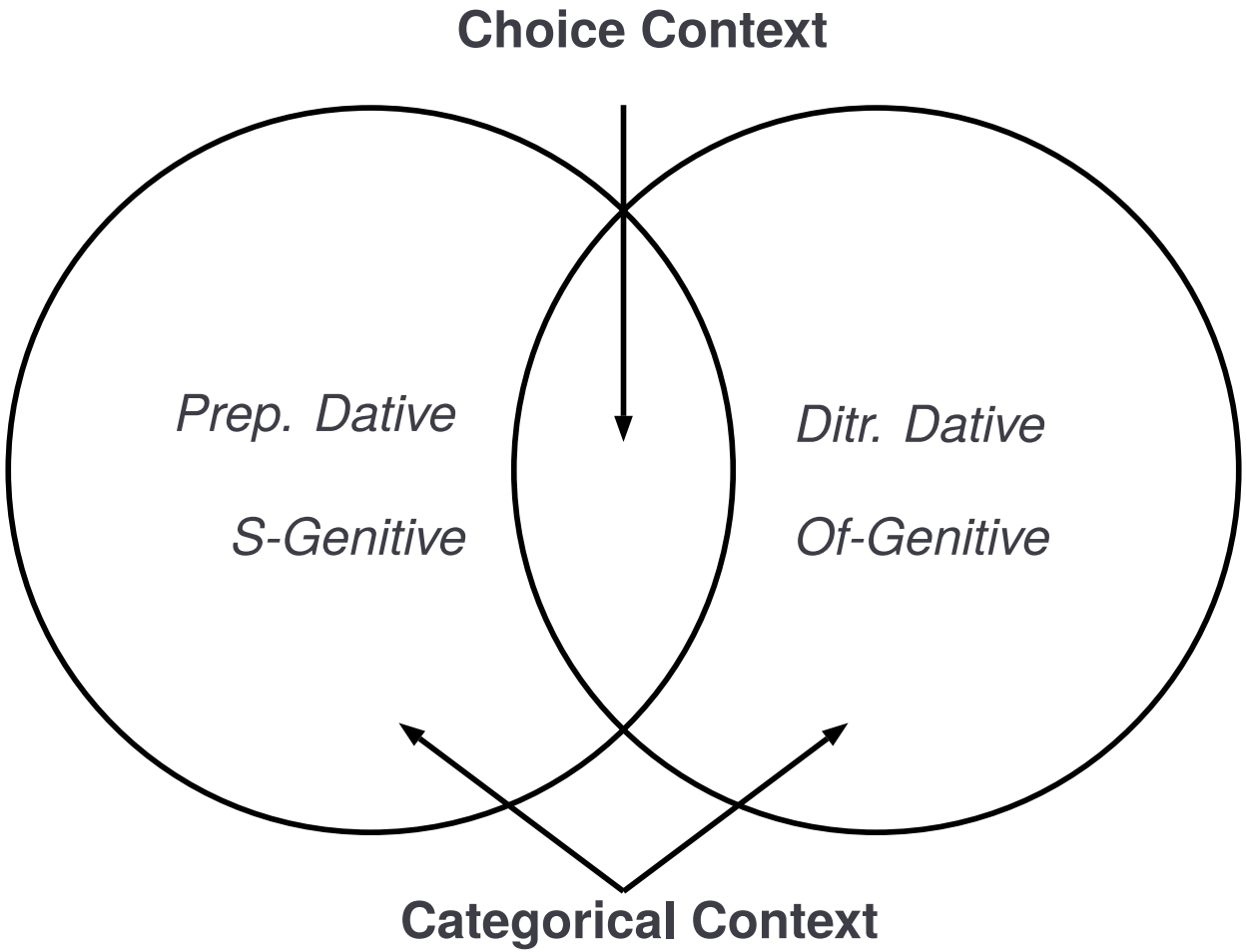


Figure: Adapted from Rosenbach (2002, 28)

DATA

CORPUS DATA


Great Britain
Ireland
Canada
Jamaica
India
Hong Kong
Philippines
Singapore
New Zealand

THE CORPUS FAMILIES

For each alternation

- 10,000 cases from ICE
- 5,000 cases from GloWbE

THE CORPUS FAMILIES

For each alternation

- 10,000 cases from ICE
- 5,000 cases from GloWbE

ICE

- Greenbaum (1996)
- 1 million words per component
- 60 % spoken, 40 % written
- traditional compilation

GloWbE

- Davies and Fuchs (2015)
- 1.9 billion words in total
- 60 % blogs (informal), 40 % other (more formal)
- large-scale

RETRIEVING DATIVE TOKENS

1. Extracting dative tokens using a verb list
2. Weeding out instances that are not interchangeable, e.g.
    - Fixed and idiomatic expressions (e.g. *bring it to the boil*)
    - Spatial goals (e.g. *send their daughter to school*)
    - Beneficiaries (e.g. *We get them uh typed photo copies*)

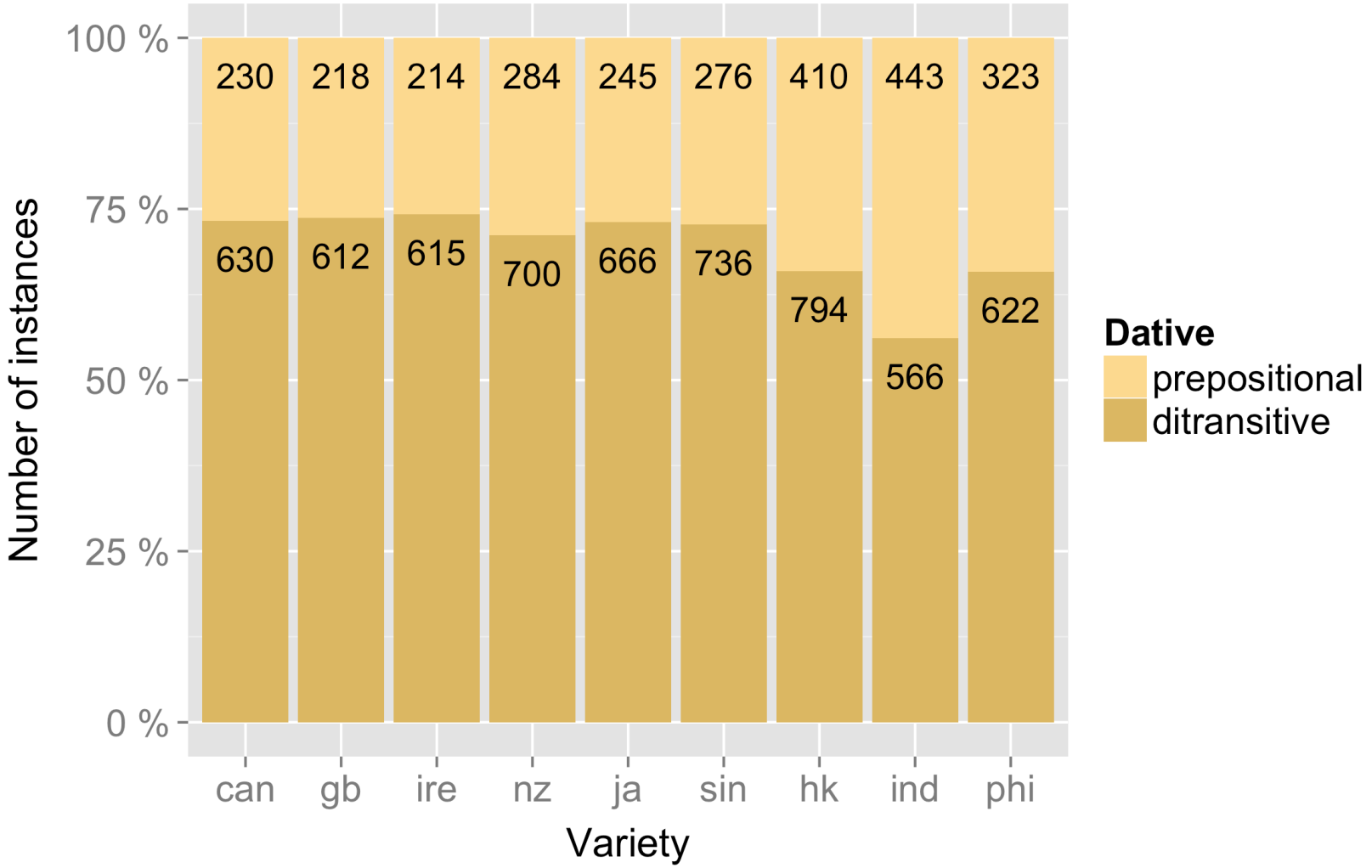
(see Bresnan et al. 2007; Bresnan and Hay 2008b; Bresnan and Ford 2010b; Bernaisch et al. 2014; De Cuypere and Verbeke 2013)

RETRIEVING GENITIVE TOKENS

1. Extract genitive tokens using the markers *'s*, *s'*, and *of*
2. Weeding out instances that are not interchangeable, e.g.
    - Indefinite *of*-constructions (e.g. *a supply of fuel*)
    - Fixed expressions (e.g. *The Bank of England*)
    - Partitive constructions (e.g. *one of my friends*)

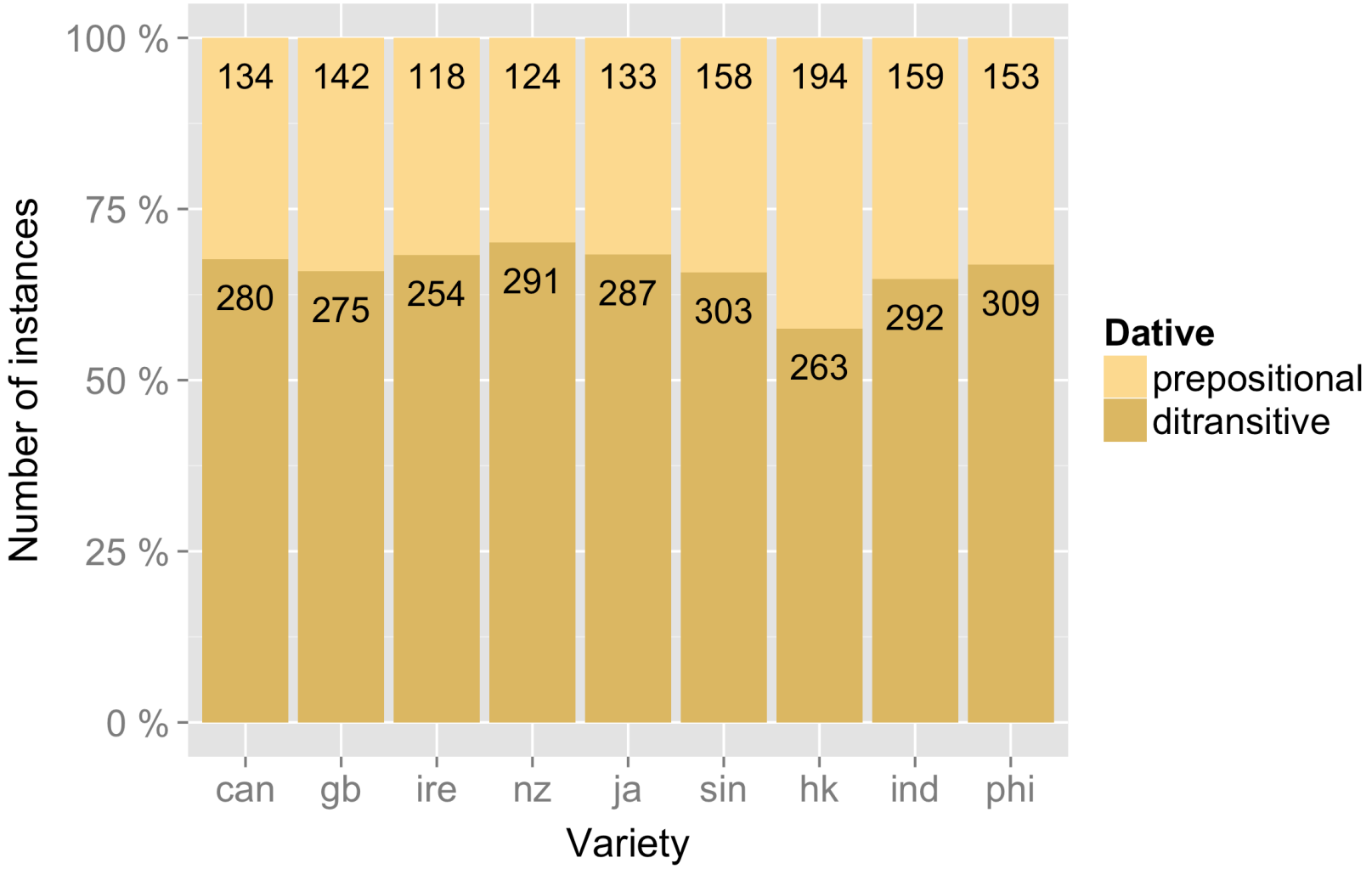
(see, e.g., Rosenbach 2002, 2014)

RAW FREQUENCIES—DATIVES BY VARIETY (ICE)



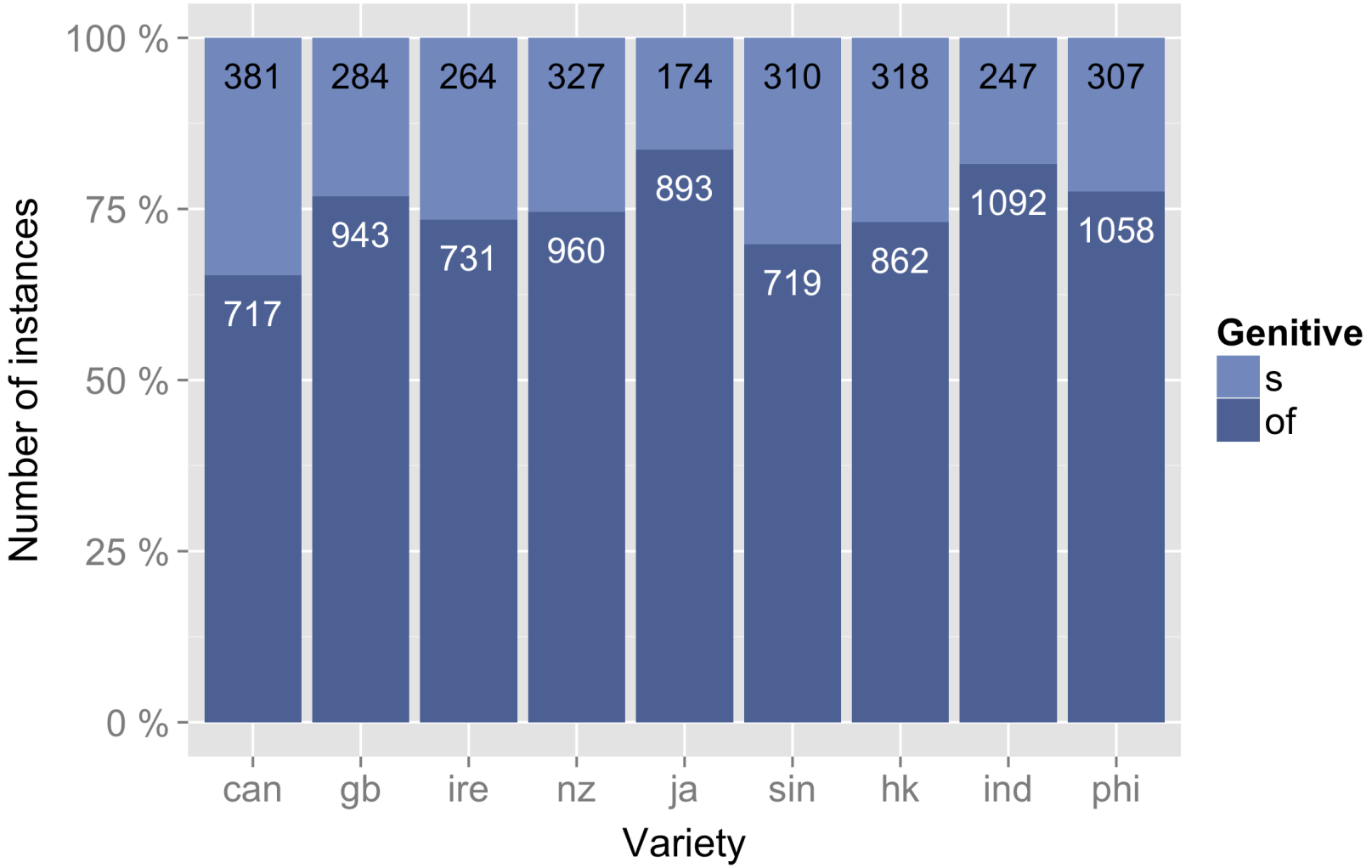
$(n = 8,548)$

RAW FREQUENCIES—DATIVES BY VARIETY (GLOWBE)



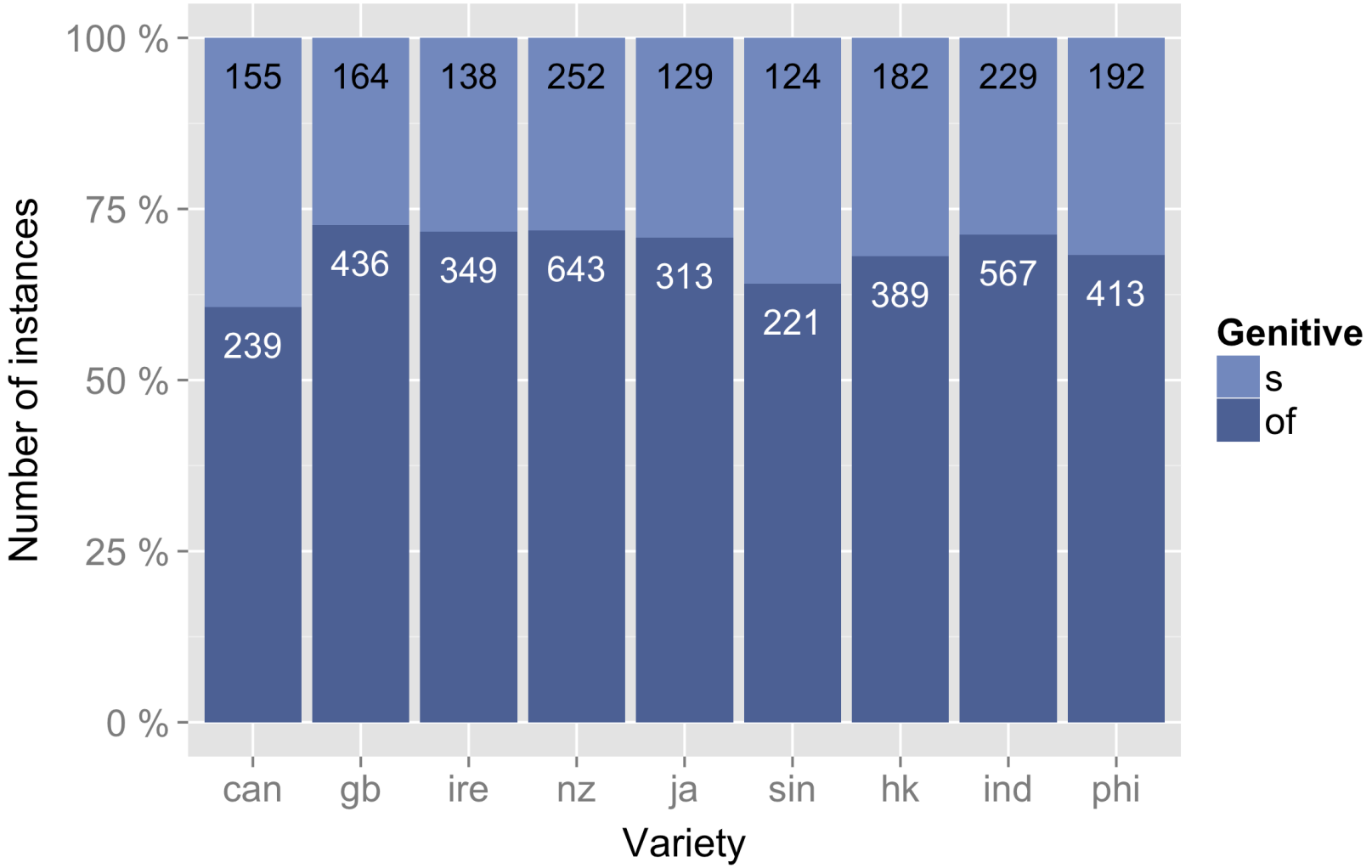
$(n = 3,869)$

RAW FREQUENCIES—GENITIVES BY VARIETY (ICE)



$(n = 10,587)$

RAW FREQUENCIES—GENITIVES BY VARIETY (GLOWBE)



$(n = 5,135)$

RAW FREQUENCIES

- Datives
  - Higher overall ratio of prepositional datives in GloWbE
  - Ratio of prepositional and ditransitive datives seems more stable

RAW FREQUENCIES

- Datives
  - Higher overall ratio of prepositional datives in GloWbE
  - Ratio of prepositional and ditransitive datives seems more stable
- Genitives
  - Higher overall ratio of *s*-genitives in GloWbE
  - Ratio of *s*-genitives and *of*-genitives seems more stable

RAW FREQUENCIES

- Datives
  - Higher overall ratio of prepositional datives in GloWbE
  - Ratio of prepositional and ditransitive datives seems more stable
- Genitives
  - Higher overall ratio of *s*-genitives in GloWbE
  - Ratio of *s*-genitives and *of*-genitives seems more stable
- Both alternations
  - The low-frequency form is comparatively frequent in GloWbE
  - Despite the smaller sample, ratios seem more stable across varieties in GloWbE

METHODOLOGY

COMMON PREDICTORS

- Animacy of the constituents (*inanimate* or *animate*)
  - Datives: recipient and theme
  - Genitives: possessor and possessum
- Syntactic weight measured as Weight Ratio
  - Dative alternation: $WR = \frac{recipient\ \ length}{theme\ \ length}$
  - Genitive alternation: $WR = \frac{possessor\ \ length}{possessum\ \ length}$
- Givenness (*given* or *new*)
  - Dative alternation: recipient and theme
  - Genitive alternation: possessor
- Corpus (*ICE* or *GloWbE*)
- Variety (*gb*, *can*, *nz*, *ire*, *ind*, *hk*, *sin*, *ja* or *phi*)

DATIVE-SPECIFIC PREDICTORS

- Person of recipient (*local* or *non-local*)
- Complexity of recipient and theme (*complex*, i.e. following postmodification, or *simple*)
- Pronominality of recipient and theme (*pron* or *non-pron*)
- Definiteness of recipient and theme (*definite* or *indefinite*)
- Concreteness of theme (*concrete*, i.e. perceivable with senses, or *non-concrete*)

GENITIVE-SPECIFIC PREDICTORS

- Final Sibilancy of the possessor
  (5) The paradox's conclusion <ICE-IND:W2B-021>
  (6) the church's solidarity with women <ICE-NZ:S1B-011>
- Thematiciy of possessor head
- Type-Token Ratio of immediate context
- Overall Frequency of the possessor head

MIXED-EFFECTS LOGISTIC REGRESSION

- Random effects (REs)
  - Predictors that are special to the sample, i.e not repeatable (Baayen, 2008, 241)
  - It is crucial to account for idiosyncrasies of speakers and corpus structure (Gries, 2015)
  - Varying intercepts for speakers and for constituent heads for both alternations
- Model selection according to guidelines in Zuur et al. (2009, ch. 5) and Gries (2015)

RESULTS

DATIVES IN GLOWBE & ICE

Main effects

Fixed effects	Estimate	Odds Ratio	p-value	
THEME_COMPLEXITYsimple	-0.63	0.53	<6.29e-07	***
REC_DEFINITENESSdef	0.52	1.68	5.60e-06	***
REC_COMPLEXITYsimple	0.95	2.59	8.06e-09	***
REC_GIVENNESSgiven	0.24	1.27	<0.02	*
REC_ANIMACYanimate	0.80	2.23	7.54e-12	***
THEME_DEFINITENESSdef	-0.70	0.50	7.36e-12	***
zTHEME_THEMATICITY	-0.34	0.71	0.0003	***
REC_PERSONlocal	1.51	4.53	0.0004	***
THEME_PRONpron	-1.35	0.26	0.0024	**
REC_PRONpron	0.91	2.48	0.024	*
WEIGHT_RATIO	-3.37	0.03	1.33e-14	***
VARIETYhk	-0.67	0.51	0.036	*
VARIETYind	-2.61	0.07	5.88e-15	***
VARIETYphi	-0.95	0.39	0.0075	**

DATIVES IN GLOWBE & ICE

Interactions

Interactions	Estimate	Odds Ratio	p-value	
CORPUSglowbe:VARind	1.21	3.35	0.0015	**
REC_PRONpron:VARcan	0.99	2.69	0.0334	*
REC_PRONpron:VARind	1.32	3.74	0.0026	**
THEME_CONCRconcrete:VARind	0.94	2.56	0.0408	*
WEIGHT_RATIO:VARind	1.26	3.53	0.0245	*

- Adding 'corpus' as an interaction term with all other predictors in this table doesn't add anything to the model.
- Accuracy: 93.5 %; C-Value: 0.98

GENITIVES IN GLOWBE & ICE

Main effects

Fixed Effect	Estimate	Odds Ratio	p-value	
POR_ANIMACYa	3.454	31.627	0.000	***
PUM_ANIMACYa	0.231	1.260	0.042	*
WR_LOG	-1.48	0.228	0.000	***
FINAL_SIBILANCY_PORtrue	-1.135	0.321	0.000	***
POR_GIVENNESSgiven	0.182	1.200	0.021	*
POR_HEAD_FREQ_LOG	-0.137	0.872	0.000	***
POR_THEMATICITY_LOG	0.22	1.246	0.000	***
TTR	2.482	11.965	0.000	***
CORPUSGloWbE	0.288	1.334	0.014	*
VARIETYcan	0.786	2.195	0.009	**
VARIETYire	0.433	1.542	0.164	*
VARIETYnz	0.593	1.809	0.043	*
VARIETYsin	0.67	1.954	0.023	*

GENITIVES IN GLOWBE & ICE

Interactions

Fixed Effects	Estimate	Odds R.	p-value	
POR_ANIMACYa:VARIETYphi	-1.158	0.314	0	***
POR_FREQ:CORPUSGloWbE	0.074	1.077	0.008	**
WR_LOG:VARIETYind	0.445	1.560	0.013	*
WR_LOG:CORPUSGloWbE	-0.215	0.807	0.015	*
FINAL_SIBILANCYtrue:VARIETYcan	-0.871	0.419	0.026	*
POR_ANIMACYa:VARIETYind	-0.638	0.528	0.041	*

- Accuracy: 91.9 %; C-Value: 0.97

REGRESSION SUMMARY

Significant main effects

	ICE	GloWbE	ICE & GloWbE
Datives	12	11	14
Genitives	10	6	12

Significant interactions with VARIETY

	ICE	GloWbE	ICE & GloWbE
Datives	3	0	4
Genitives	6	1	4

REGRESSION SUMMARY II

- Datives
  - Two more main effects and one interaction reached significance.
  - Only Indian English significantly different in GloWbE.

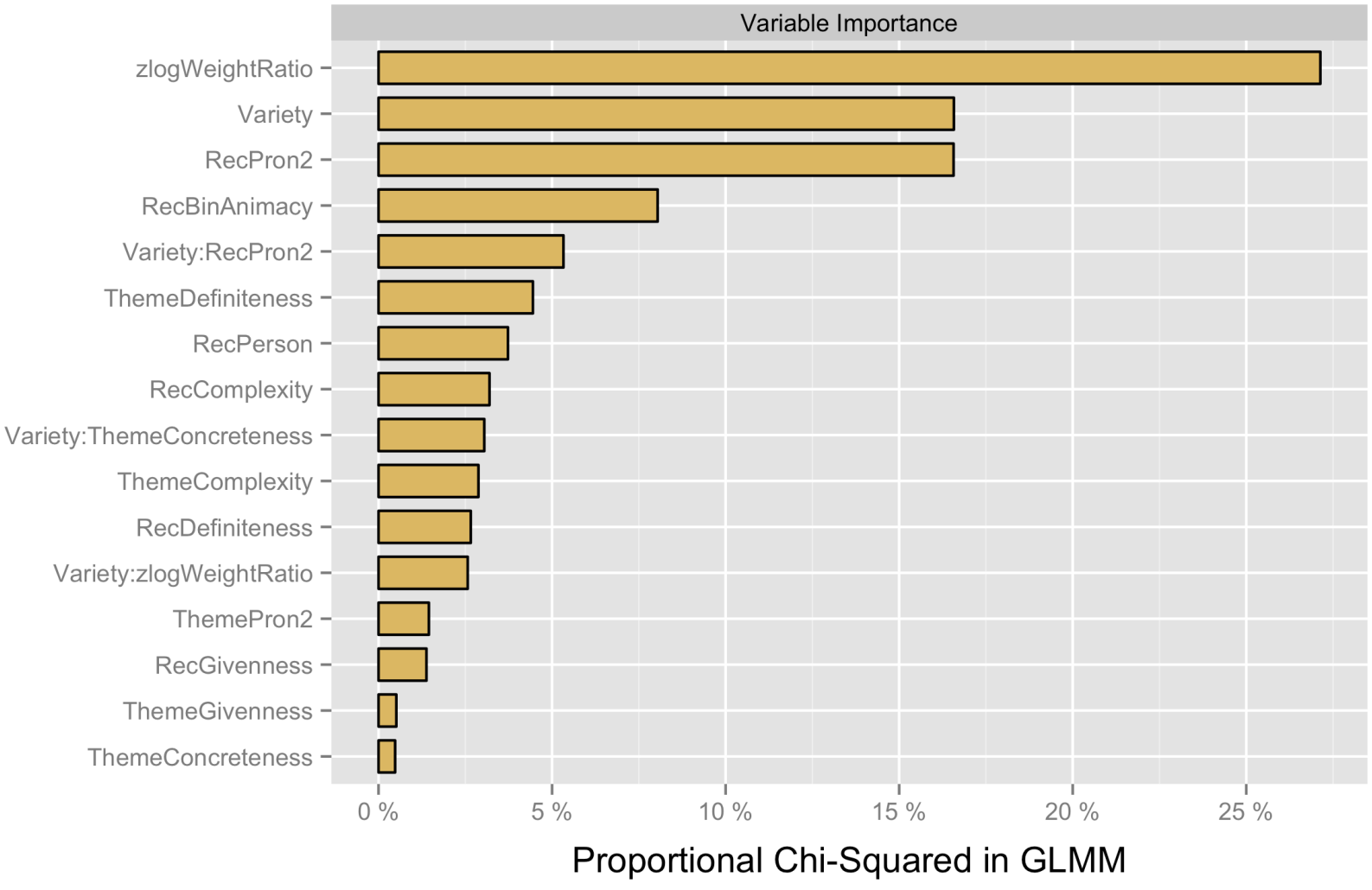
REGRESSION SUMMARY II

- Datives
  - Two more main effects and one interaction reached significance.
  - Only Indian English significantly different in GloWbE.
- Genitives
  - Two additional main effects reached significance, but two interaction terms were lost.
  - Weight Ratio more influential in GloWbE.
  - Frequent possessor heads less averse to *s*-genitive in GloWbE.
  - *s*-genitives slightly more in entire GloWbE sample.
  - No significant Corpus-Variety interaction

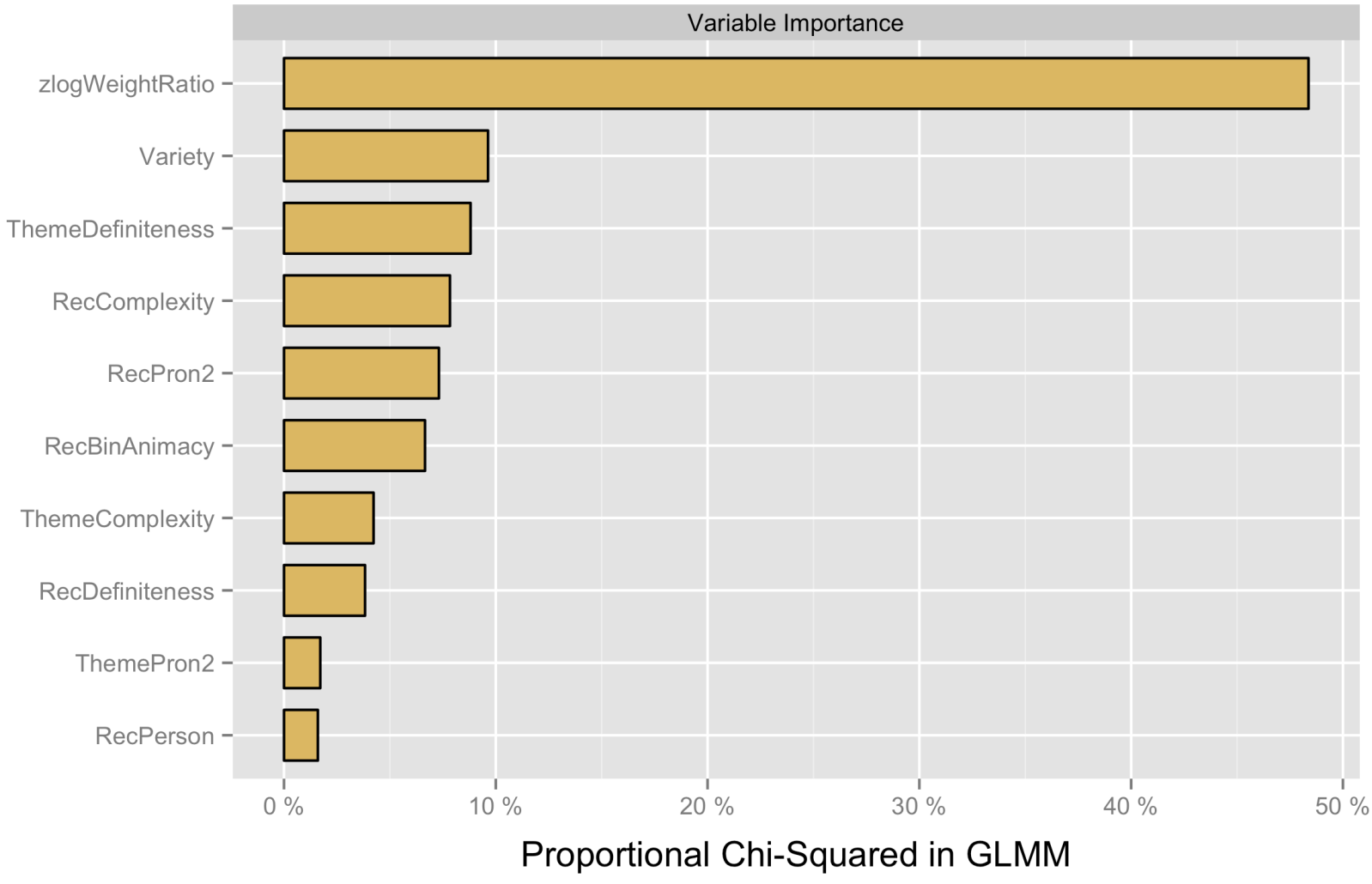
REGRESSION SUMMARY II

- Datives
  - Two more main effects and one interaction reached significance.
  - Only Indian English significantly different in GloWbE.
- Genitives
  - Two additional main effects reached significance, but two interaction terms were lost.
  - Weight Ratio more influential in GloWbE.
  - Frequent possessor heads less averse to *s*-genitive in GloWbE.
  - *s*-genitives slightly more in entire GloWbE sample.
  - No significant Corpus-Variety interaction
- Both alternations
  - No surprises (e.g. change of direction of an effect) when comparing the ICE and the GloWbE models.
  - No additional interactions between Corpus and any other predictor in both models.

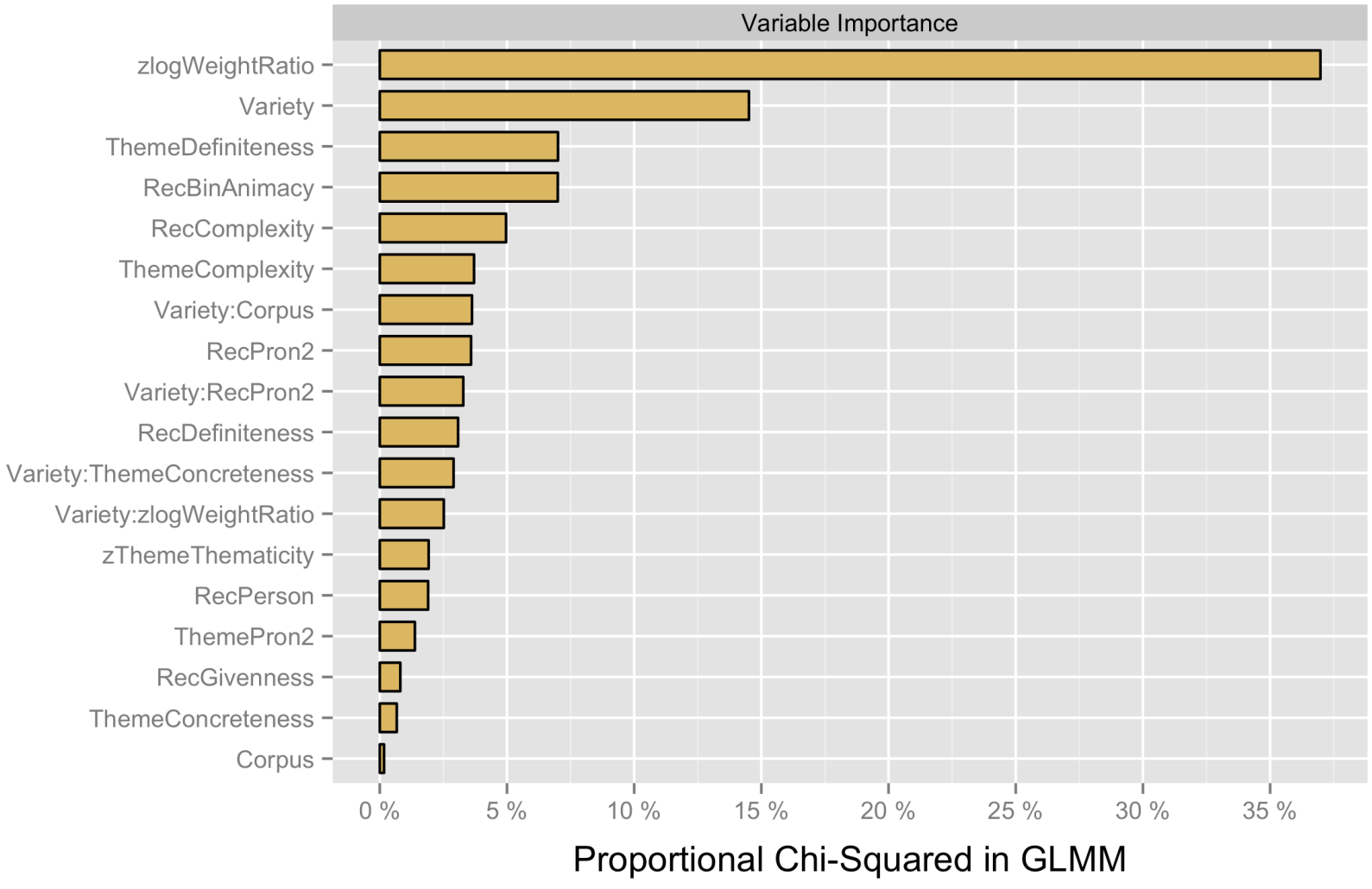
VARIABLE IMPORTANCE (ICE): DATIVES



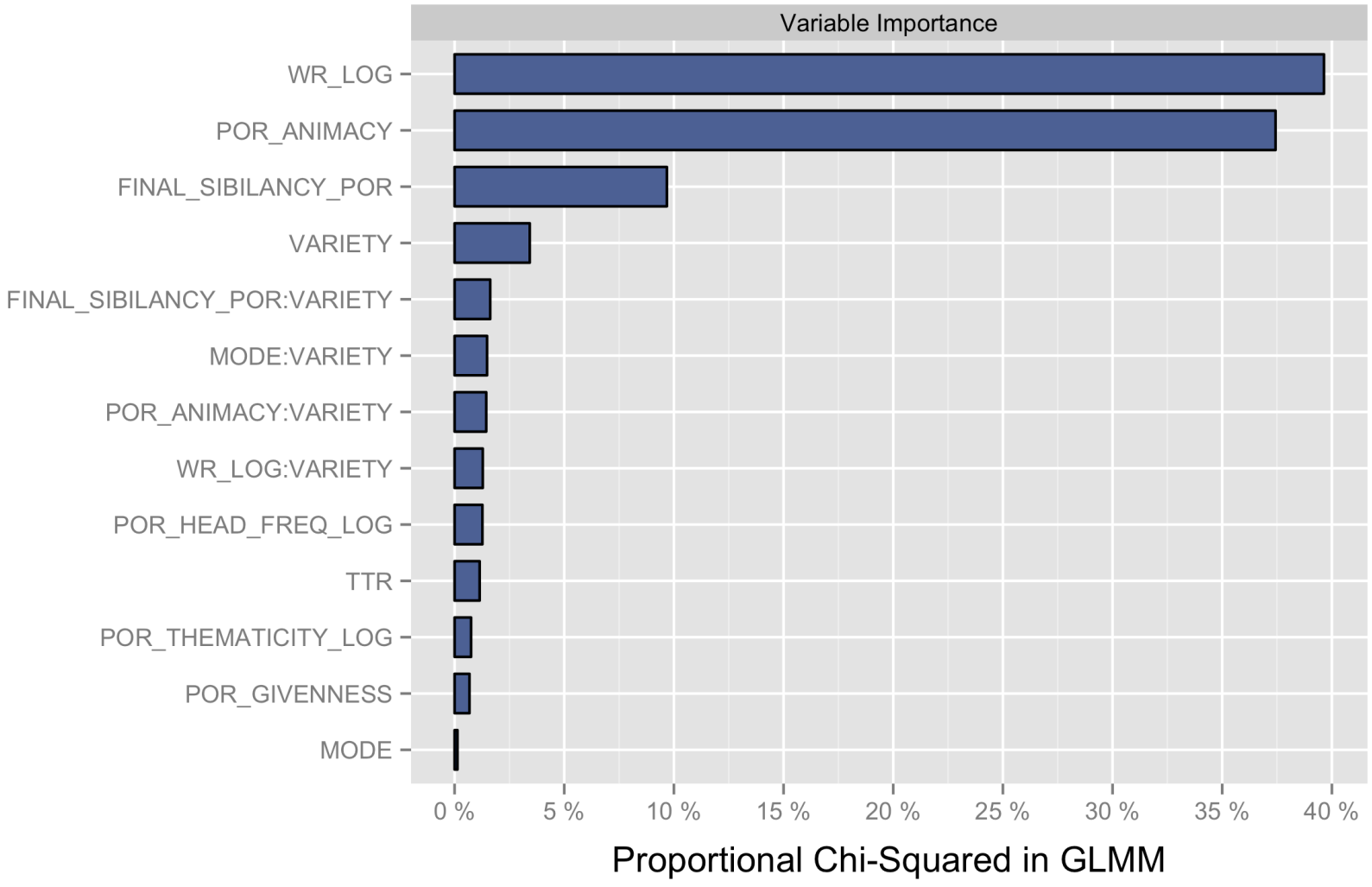
VARIABLE IMPORTANCE (GLOWBE): DATIVES



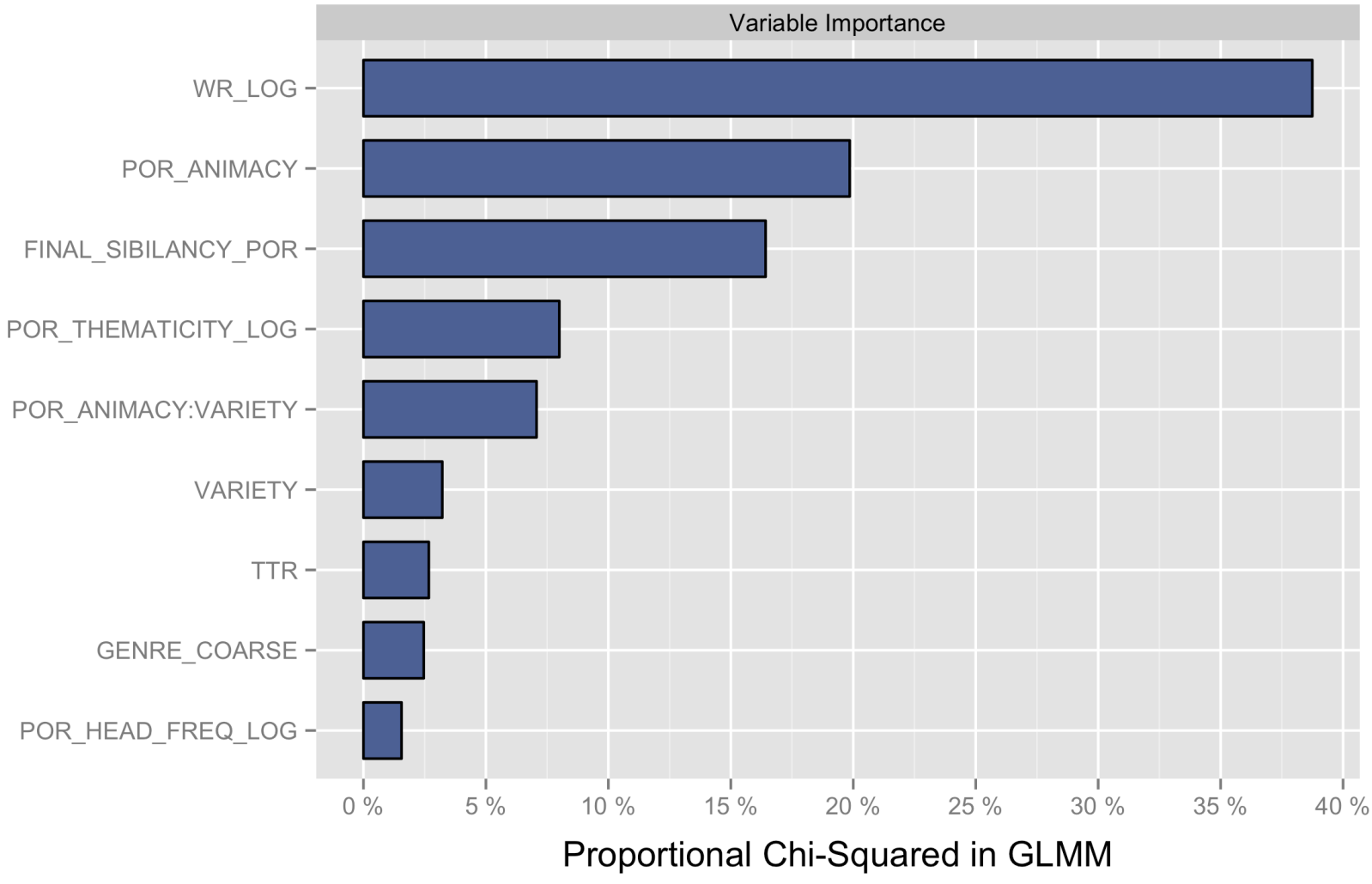
VARIABLE IMPORTANCE (FULL DATASET): DATIVES



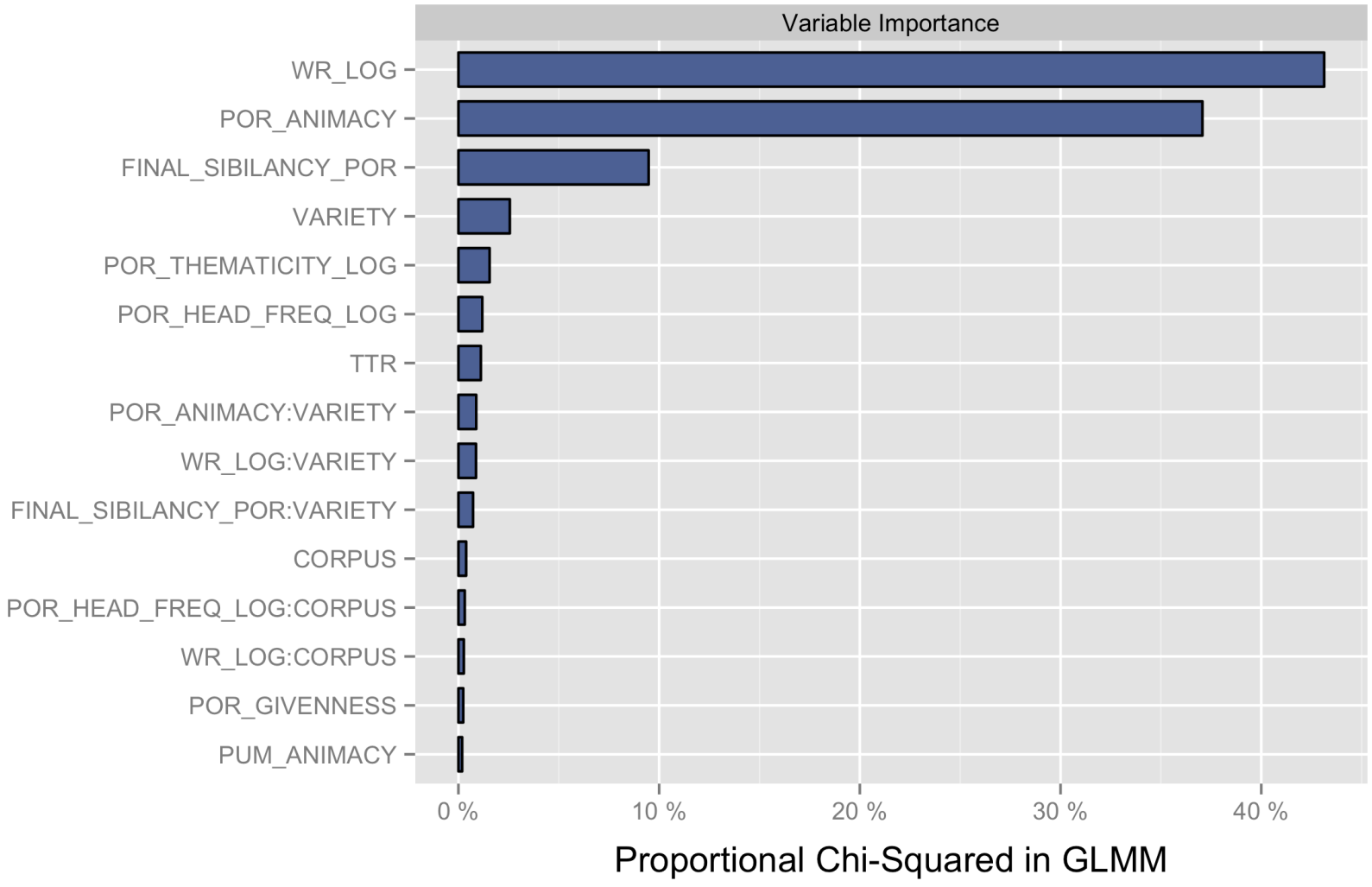
VARIABLE IMPORTANCE (ICE): GENITIVES



VARIABLE IMPORTANCE (GLOWBE): GENITIVES



VARIABLE IMPORTANCE (FULL DATASET): GENITIVES



VARIABLE IMPORTANCE

Summary

- Datives
  - Variety lost importance slightly.
  - Recipient Pronominality (rank 2 in ICE) becomes less important (rank 8).
  - Theme Definiteness becomes more important (rank 6 to 3).
- Genitives
  - No changes.

SUMMARY & CONCLUSION

SUMMARY & CONCLUSION

- Discovering cross-varietal differences in the mechanisms of the dative and the genitive alternation requires a lot of richly-annotated data.
- The supplementation of data form GloWbE turned out to be beneficial for our research purposes
- There is an amazing similarity between ICE and GloWbE in the patterning of the alternations: CORPUS does not really matter
  - We discovered more significant effects in the full models.
  - Just one CORPUS-VARIETY interaction in the dative model; none in the genitive model.
  - Just slight differences in variable importance for the datives; no differences for the genitives.

REFERENCES

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Bernaisch, T., S. T. Gries, and J. Mukherjee (2014). The dative alternation in South Asian English(es): Modelling predictors and predicting prototypes. *English World-Wide 35*(1), 7–31.

Bresnan, J. (2007). Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In S. Featherston and W. Sternefeld (Eds.), *Roots: Linguistics in Search of Its Evidential Base*, pp. 75–96. Berlin: Mouton de Gruyter.

Bresnan, J., A. Cueni, T. Nikitina, and B. Harald (2007). Predicting the Dative Alternation. In G. Boume, I. Kraemer, and J. Zwarts (Eds.), *Cognitive Foundations of Interpretation*, pp. 69–94. Amsterdam: Royal Netherlands Academy of Science.

Bresnan, J. and M. Ford (2010a). Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language 86*(1), 168–213.

Bresnan, J. and M. Ford (2010b). Predicting Syntax: Processing dative constructions in American and Australian Varieties of English. *Language 86*(1), 168–213.

Bresnan, J. and J. Hay (2008a, February). Gradient grammar: An effect of animacy on the syntax of give in New Zealand and American English. *Lingua 118*(2), 245–259.

Bresnan, J. and J. Hay (2008b, February). Gradient grammar: An effect of animacy on the syntax of give in New Zealand and American English. *Lingua 118*(2), 245–259.

Davies, M. and R. Fuchs (2015). Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE). *English World-Wide 36*(1), 1–28.

De Cuypere, L. and S. Verbeke (2013, June). Dative alternation in Indian English: A corpus-based analysis. *World Englishes 32*(2), 169–184.

Gahl, S. and S. Garnsey (2004). Knowledge of Grammar, Knowledge of Usage: Syntactic Probabilities Affect Pronunciation Variation. *Language 80*, 748–775.

Gahl, S. and A. C. Yu (2006). *Special theme issue: Exemplar-based models in linguistics*. The linguistic review. Mouton de Gruyter.

Greenbaum, S. (1996). Introducing ICE. In *Comparing English Worldwide: The International Corpus of English*, pp. 3–12. Oxford: Clarendon Press.

Gries, S. T. (2015). The most under-used statistical method in corpus linguistics: multi-level (and mixed-effects) models. *Corpora 10*(1), 95–125.

Kachru, B. B. (Ed.) (1992). *The Other tongue: English across cultures* (2nd ed ed.). English in the global context. Urbana: University of Illinois Press.

Labov, W. (1982). Building on empirical foundations. In W. Lehmann and Y. Malkiel (Eds.), *Perspectives on Historical Linguistics*, pp. 17–92. Amsterdam, Philadelphia: Benjamins.

Mesthrie, R. and R. M. Bhatt (2008). *World Englishes: the study of new linguistic varieties*. Key topics in sociolinguistics. Cambridge, UK ; New York: Cambridge University Press.

Rosenbach, A. (2002). *Genitive Variation in English. Conceptual Factors in Synchronic and Diachronic Studies*. Mouton de Gruyter.

Rosenbach, A. (2014, July). English genitive variation – the state of the art. *English Language and Linguistics 18*(02), 215–262.

Schneider, E. (2007). *Postcolonial English: Varieties Around the World*. Cambridge University Press.

Szmrecsanyi, B., J. Grafmiller, B. Heller, and M. Röthlisberger (2016). Around the world in three alternations: Modeling syntactic variation in global varieties of english. *English World-Wide 37*(2).

Wolk, C., J. Bresnan, A. Rosenbach, and B. Szmrecsanyi (2013). Dative and genitive variability in Late Modern English: Exploring cross-constructional variation and change. *Diachronica 30*(3), 382–419.

Zuur, A. F., E. N. Ieno, N. J. Walker, A. A. Saveliev, and G. M. Smith (2009). *Mixed Effects Models and Extensions in Ecology with R*. Springer.